MODULATION FREQUENCY FEATURES FOR AUDIO FINGERPRINTING

Somsak Sukittanon and Les E. Atlas
Department of Electrical Engineering, University of Washington, Box 352500, Seattle, Washington 98195-2500, USA
{ssukitta, atlas}@ee.washington.edu

ABSTRACT

This paper explores modulation frequency features with subband normalization for audio identification. Our main goal is to find features for audio fingerprinting that are invariant to time and frequency distortions, both unintentional and intentional. Two-dimensional features, called "joint acoustic and modulation frequency," are proposed. The paper describes these features and corresponding cross entropy classification. Experimental results show that standard spectral features are inadequate when frequency distortion occurs, as in low bit rate coding or equalization. In contrast, our proposed normalized modulation frequency features can provide accurate fingerprints, even when time and frequency distortions are imposed on music passages.

1. INTRODUCTION

Audio fingerprint technology has been proposed for reporting property rights or monitoring radio broadcasts [1]. In this approach, a computer or other electronic device that has access to a central database of statistics automatically labels unlabeled music. An accurate fingerprint can ideally identify audio content based solely on acoustic properties. Moreover, each song needs to be represented compactly. Previous research into acoustic feature-based audio retrieval has focused on using short-term spectral estimates and related features [2]. While these results were quite promising, they also showed that there was no universal feature that gives high recognition rates under both time and frequency distortions. The goal of this paper is to present a new feature type that maintains excellent labeling (classification) accuracy under common unintentional and intentional distortions.

Modulation frequency analysis is proposed to characterize the time-varying behavior of the audio signal. As illustrated in Fig. 1, the audio signal is first decomposed into subbands using Fourier analysis. We call the standard spectral density features that result from this first step "acoustic frequency." Then a modulation frequency transform is performed on each subband to provide joint acoustic and modulation frequency features.

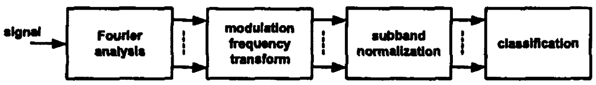

Figure 1: Joint acoustic and modulation frequency analysis.

This joint acoustic and modulation frequency representation is a transform in time of a demodulated (e.g. magnitude-squared) short-time spectral estimate. For most signals, such as music, short-time spectral estimates change with time. If a signal is observed for a long period of time and the spectrum changes periodically, then the signal can be modeled as a cyclostationary signal [3]. Similar joint frequency analysis, without the assumption of a long time estimate, has also been recently applied to audio coding [4].

Audio passages supplied to a fingerprinting system can be altered by linear or nonlinear transmission or encoding/decoding systems. One common example of linear distortion is frequency equalization, which amplifies or attenuates frequency subbands. The effects of this frequency distortion on the above joint frequency features are predictable and removable; for each acoustic frequency subband the amplitude of all modulation frequency features is scaled uniformly. Similar behavior exists for common nonlinear distortions. Feature normalization after joint acoustic and modulation frequency feature analysis is thus proposed and shown to compensate for audio channel distortions.

2. JOINT ACOUSTIC AND MODULATION FREQUENCY REPRESENTATION

One possible joint acoustic and modulation frequency representation, $P(\eta,\omega)$, is a two-dimensional transform of the instantaneous autocorrelation function of the signal (2.1) where τ, ω, and η are delayed time, acoustic frequency, and modulation frequency, respectively.

$$P(\eta,\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^{*}(t-\tau/2)x(t+\tau/2)e^{-j\omega\tau}e^{-j\eta t}\,d\tau dt \qquad (2.1)$$

A simple amplitude modulation (AM) model (2.2), where ω_m and ω_c are the modulation and carrier frequencies, respectively, can be used to illustrate behavior on this joint acoustic and modulation frequency plane. For this AM model, $P(\eta,\omega)$ should ideally have non-zero Dirac impulse terms at only $P(0,\pm\omega_c)$ and $P(\pm\omega_m,\pm\omega_c)$. As seen in (2.3), where $*_{\eta,\omega}$ denotes convolution in η and ω, and Fig. 2a, there are also non-zero terms occurring at much higher modulation frequencies, $\eta=\pm 2\omega_c$, and at double modulation frequencies, $\eta=\pm 2\omega_m$. This last term can be interpreted as interference due to the quadratic nature of (2.1).

$$x(t) = \left(1+\cos\omega_m t\right)\cos\omega_c t \qquad (2.2)$$

$$\begin{aligned}P(\eta,\omega) = {} & \pi/8\{4\delta(\eta,\omega)+2\delta(\eta\pm\omega_m,\omega\pm\omega_m/2)+\delta(\eta,\omega\pm\omega_m)+ \\ & \delta(\eta\pm 2\omega_m,\omega)\}*_{\eta,\omega}\{\delta(\eta\pm 2\omega_c,\omega)+\delta(\eta,\omega\pm\omega_c)\} \qquad (2.3)\end{aligned}$$

The non-ideal effects seen in (2.3), and thus present in most arbitrary signals, can be significantly reduced by the inherent smoothing properties of the spectrogram. Referring to Fig. 1, if the first Fourier analysis step is simply estimated by the magnitude square of a short time Fourier transform with an appropriately chosen window length and the second transform also uses Fourier bases, the above non-ideal terms are smoothed and attenuated. The result for this more practical case is illustrated in Fig. 2b.

0-7803-7402-9/02/$17.00 ©2002 IEEE

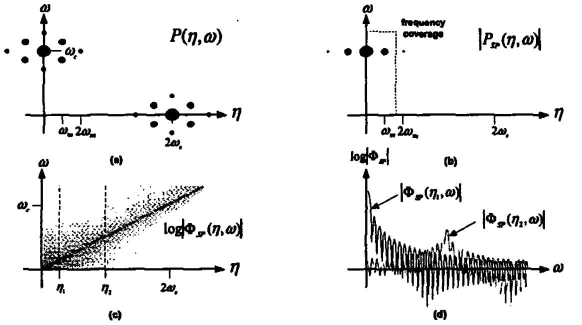

Figure 2: (a) Joint frequency representation of an AM signal, (b) a smoothed representation using a time transform of a spectrogram, (c) joint representation of a rectangular window, and (d) $\Phi_{SP}(\eta,\omega)$ for a fixed low (η_1) and high (η_2) modulation frequency.

To proceed with more general analysis, the instantaneous autocorrelation function of any bilinear joint time-frequency representations, e.g. a spectrogram, can be written in a general form [5]. The instantaneous autocorrelation function of spectrogram, $R_{SP}(t,\tau)$, can be viewed as the smoothed instantaneous autocorrelation of the signal where the smoothing kernel, $\phi_{SP}(t,\tau)$, is convolutional in t and multiplicative in τ (2.4). Since the joint frequency representation is the two-dimensional transform of the instantaneous autocorrelation function, $P_{SP}(\eta,\omega)$ can be viewed as the smoothed $P(\eta,\omega)$. The transform of the smoothing kernel, $\Phi_{SP}(\eta,\omega)$, is the joint frequency representation of the spectrogram window, $h(t)$ (i.e. a rectangular or tapered window function), and is convolutional in ω and multiplicative in η, (2.5)-(2.6). For a rectangular window, the transform kernel $\Phi_{SP}(\eta,\omega)$ has a $\sin x/x$ behavior along the heaviest diagonal line of Fig. 2c. As shown in Fig. 2d, the magnitude of the mainlobe of $\Phi_{SP}(\eta,\omega)$ decreases with increasing η.

$$R_{SP}(t,\tau) = \phi_{SP}(t,\tau) *_t x^*(t-\tau/2)x(t+\tau/2) \tag{2.4}$$

$$\begin{aligned} P_{SP}(\eta,\omega) &= (1/2\pi)\iint R_{SP}(t,\tau)e^{-j\omega\tau}e^{-j\eta t}d\tau dt \\ &= \Phi_{SP}(\eta,\omega) *_\omega P(\eta,\omega) \end{aligned} \tag{2.5}$$

$$\Phi_{SP}(\eta,\omega) = \iint h^*(t-\tau/2)h(t+\tau/2)e^{-j\omega\tau}e^{-j\eta t}d\tau dt \tag{2.6}$$

As implied by the change from Fig. 2a to 2b, the sidelobes of $\Phi_{SP}(\eta,\omega)$ essentially remove the non-zero terms around $\eta=\pm2\omega_c$ and the convolution in the ω direction causes the terms at $\eta=0$ and $\pm\omega_m$ to be merged vertically. For our application only positive and low modulation frequencies are needed. The dotted line in Fig. 2b shows the coverage of the feature plane used in this paper. $P(0,\omega)$ represents a long-term power spectral estimate of the signal. Features lying on the plane where $\eta>0$ are modulation frequency estimates.

Recent psychoacoustic results [6] suggest that a log frequency scale, with resolution consistent with a constant-Q over the whole range, best mimics human perception of modulation frequency. Our approach uses a continuous wavelet transform to efficiently approximate this constant-Q effect. For discrete scales, s, the wavelet filter, $\psi(t)$, is applied along each temporal row of the spectrogram output. A sum across the wavelet translation axis, ζ, is performed to produce a joint frequency representation with nonuniform frequency resolution on the modulation frequency axis, now called the s axis, (2.7)-(2.9).

$$P_{SP}(t,\omega) = (1/2\pi)\left|\int x(u)h^*(u-t)e^{-j\omega u}du\right|^2 \tag{2.7}$$

$$P_{SP}(s,\zeta,\omega) = \int P_{SP}(t,\omega)\psi^*(t-\zeta/s)dt \tag{2.8}$$

$$P_{SP}(s,\omega) = \int\left|P_{SP}(s,\zeta,\omega)\right|^2 d\zeta \tag{2.9}$$

Fig. 3 shows an advantage of this constant-Q bandwidth. An AM signal, with fixed carrier frequency of 2756 Hz and two modulation frequencies, 3 and 24 Hz (loosely corresponding to modulation rates seen in music), was simulated and is shown with Fourier and constant-Q modulation frequency decompositions. Both cases have the same total number of features. Ideally, the modulation frequency dimension should show separate regions representing the two distinct modulation frequencies. With the same total dimensionality, the Fourier modulation decomposition cannot resolve these modulation frequencies.

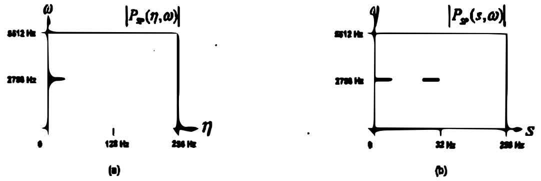

Figure 3: Joint frequency analysis of a two-component AM signal (a) using a Fourier modulation decomposition and (b) constant-Q modulation decomposition.

A discrete implementation is shown in Fig. 4. $P[s,k]$ is the discrete version of $P_{SP}(s,\omega)$ where s and k are dyadic wavelet scale and discrete acoustic frequency, respectively. Consistent with the constant-Q resolution, the scale variable is inversely proportional to the logarithm of modulation frequency in Hertz. We assume the signal is band-limited, with f_m^{max} and f_c^{max} as the highest modulation and carrier frequencies in Hz, respectively. First, the digitized signal, with sampling rate F_s Hz, is decomposed into K discrete acoustic frequencies. A first decimation step, D_1, reduces the amount of data in each channel. The lower bound for D_1, (2.10), avoids the possibility of subsequent aliasing of the subband signals. A magnitude-square envelope detector then demodulates each subband. (Other demodulation techniques, such as Hilbert and cepstral, could also be used.) The second decimation step, D_2, is applied to further reduce the final feature size. D_2 depends on D_1 and f_m^{max}, (2.11).

$$D_1 \le F_s/4f_c^{max} \tag{2.10}$$

$$D_2 \le F_s/4D_1f_m^{max} \tag{2.11}$$

Continuing to work through Fig. 4, a tapered window is used to reduce the sidelobes of the subsequent modulation frequency estimate. Modulation frequency is then estimated for each subband independently. Finally, the output features include modulation frequency features, $P_{mod}[s,k]$, and long-term spectral estimate features, $P_{spec}[k]$.

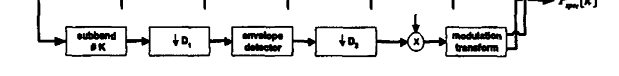

Figure 4: Joint frequency feature extraction. Details of parameter choices are in Section 4.3.

3. CROSS ENTROPY CLASSIFICATION

Referring to Fig. 1, classification of a test signal was performed after feature extraction. Due to computational advantages of a simple centroid computation, a cross entropy approach was used. This approach also fits the sparseness of joint frequency features. A feature set representing a test song is $P=\{p_1,p_2,...p_M\}$. p_m are the joint frequency features for a time frame at time m and M is the total number of frames in an audio passage used to represent a song. For passages in the database, we define the feature set of the reference audio as $Q^i=\{q_1^i,q_2^i...q_M^i\}$ where i is the ordered index in the database and q_m are the reference features at time m. Let $\tilde{P}$ be the centroid of P. Using a cross entropy estimate (3.1)-(3.2), the distortion between $\tilde{P}$ and Q^i is equivalent to the more efficient distortion between $\tilde{P}$ and the geometric mean of Q^i.

$$d(\tilde{P},q_m^i) = \sum_{s,k}\tilde{P}[s,k]\log\frac{\tilde{P}[s,k]}{q_m^i[s,k]} \tag{3.1}$$

$$\begin{aligned} d(\tilde{P},Q^i) &= \frac{1}{M}\sum_{m,s,k}\tilde{P}[s,k]\log\frac{\tilde{P}[s,k]}{q_m^i[s,k]} \\ &= \sum_{s,k}\tilde{P}[s,k]\log\frac{\tilde{P}[s,k]}{(\prod_m q_m^i[s,k])^{1/M}} \\ &= \sum_{s,k}\tilde{P}[s,k]\log\frac{\tilde{P}[s,k]}{\tilde{Q}^i[s,k]} \\ &= d(\tilde{P},\tilde{Q}^i) \qquad \textit{where } \tilde{Q}^i = \sqrt[M]{q_1^i q_2^i ... q_M^i} \end{aligned} \tag{3.2}$$

If a tested song is exactly the same data samples as a song represented in the database, then $d(\tilde{P},\tilde{Q}^i)$ becomes zero. Given $d(\tilde{P},\tilde{Q}^i)=0$ and $\tilde{P}[s,k]=\tilde{Q}^i[s,k]$ for all s and k implies that $\tilde{P}$ is the geometric mean of P. As shown in (3.3), for nearest neighbor classification using cross entropy, the estimation of the closest known passage in the database is done by choosing the index corresponding to the lowest distortion.

$$\begin{aligned} i^* &= \underset{i}{\mathrm{argmin}}\ d(\tilde{P},\tilde{Q}^i) \\ &= \underset{i}{\mathrm{argmax}} \sum_{s,k}\tilde{P}[s,k]\log\tilde{Q}^i[s,k] \end{aligned} \tag{3.3}$$

4. EXPERIMENTS

4.1 Data Collection

The audio data used here were recorded from commercial compact discs (CDs) with 44.1 KHz sampling rate. The database contained a wide variety of music genres with 20 second passages for each song. For our experiments, we divided the database into two sets of independent songs, drawn from independent CDs. The first set of the database was a design set, consisting of 3416 distinct songs. This set was used to empirically optimize free design parameters, such as a decision threshold. The other set was a validation set, consisting of 6570 distinct songs, used to test the accuracy of the method.

4.2 Distortions

One of the requirements for the audio fingerprint system is robustness to signal distortions. Transmission channels and users alter the recording in many ways that still leave the audio useful. Allamanche *et al* [2] suggested a list of the possible degradations that exists in practical applications. We chose a subset of those degradations. The frequency distortions used were low bit rate perceptual audio coding, MP3 and Window Media Audio® at 64 Kbps, and frequency equalization. There were 4 different types of frequency equalizers including 3 default presets (pop, rock, and classical presets) from Winamp® software and 1 preset which had extremes of attenuation from −12dB to +12dB. For time distortions, linear time shifts and dynamic time normalization, sometimes called automatic gain control (AGC), were used. Unaltered recordings were used to produce the audio fingerprint database and transformed (distorted) audio passages were used for testing.

4.3 Feature Extraction

Each song was represented by a 15-second digital audio passage. This passage was resampled to 11025 Hz and converted, via an average of the left and right channels, to mono. The resampled audio was windowed into multiple frames with a 4 second frame length and 1 second frame rate, corresponding to 12 frames for each passage. With a 4 second window, the system can estimate modulation frequencies down to 0.25 Hz or roughly 15 beats per minute. As illustrated in Fig. 4, subband filtering was performed to produce acoustic frequency using a short time Fourier transform. Since the bark frequency scale has been used in many audio applications to mimic the reduced number of channels of the human auditory system, combinations of linear acoustic frequency channels were used to reduce feature size. With the 11025 Hz sampling rate, there were 19 bark-spaced subband filters. The envelope in each subband was detected by a magnitude square operator. A lowpass filter was then used to remove the alising before the sampling rate was reduced to 512 Hz. To reduce the interference of large DC components of the subband envelope, the signal was demeaned before modulation frequency estimation. After demeaning, this modulation frequency estimation used a wavelet transform. 8 dyadic scales of biorthogonal filters were chosen to produce wavelet energy and then were summed across translation to produce a modulation frequency vector for each original subband. A scaling filter at the highest scale of the wavelet filter was used for long-term spectral features. After processing all 12 frames, a single centroid was determined via a geometric mean. The 15-second passage was thus represented by 19x9 vector of features, our proposed audio fingerprint.

4.4 Feature Choices and Parameter Design

For the use in system design, fingerprints were made for all 3416 songs in the design set. Three types of features were compared: long-term spectral estimates ($\tilde{P}_{spec}$), modulation frequency ($\tilde{P}_{mod}$), and modulation frequency with subband normalized ($\tilde{P}_{norm}$) features (4.1). As shown below, $\tilde{P}_{norm}$ can remove the effect of distortions, especially frequency equalization (EQ). Equation (4.2) shows the distortion measurement as applied to these normalized features.

$$\tilde{P}_{norm}[s,k] = \frac{\tilde{P}_{mod}[s,k]}{\sum_s \tilde{P}_{mod}[s,k]} \tag{4.1}$$

$$d(\tilde{P},\tilde{Q}^i) = \frac{1}{K}\sum_{s,k}\tilde{P}_{norm}[s,k]\log\frac{\tilde{P}_{norm}[s,k]}{\tilde{Q}^i_{norm}[s,k]} \tag{4.2}$$

Since the distortion value is always positive, a similarity value, bounded between 0 and 1, can be expressed by exponentiating the distortion score, $e^{-d(\tilde{P},\tilde{Q}^i)}$. If the similarity value between a test song and the closest song in the database exceeded a predetermined threshold then the test song was detected, otherwise the test song was considered to be an unknown song (i.e. not in the database). The predetermined threshold was chosen from a receiver operating characteristic (ROC). The similarity score between a test song and the closest song in the database (excluding the identical test song) was used to find the unknown class distribution. Fig. 5 shows the ROC before and after applying EQ. Degradation occurs significantly in $\tilde{P}_{spec}$ and $\tilde{P}_{mod}$, and slightly in $\tilde{P}_{norm}$. The threshold was selected from $\tilde{P}_{norm}$ with average accuracy of 98.3% and a false alarm rate of 2.3%. The performance of three features with a one second time misalignment is shown in Table 1.

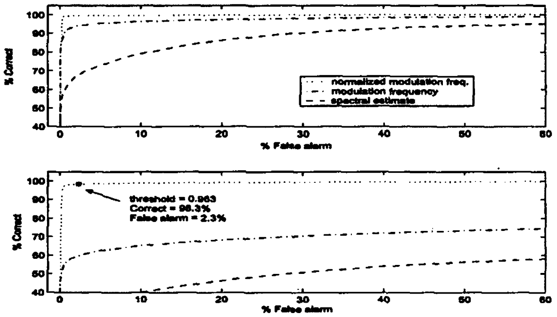

Figure 5: ROC plot of a one second time misalignment, without frequency EQ on the top and with EQ on the bottom plot.

Table 1: Performance of various feature sets on the different distortions with a one second time misalignment.

Features	Distortions		
	Compression	*Equalization*	*Normalization*
spectral estimate	90.5 %	25.2 %	88.4 %
modulation frequency	91.3 %	30.9 %	87.7 %
normalized modulation freq.	99.5 %	97.3 %	99.5 %

4.5 Results

The validation set of 6570 songs was used to predict the accuracy of the system. Only the best performing feature set, $\tilde{P}_{norm}$, was used for these validation experiments. A two second time misalignment and a uniform randomly chosen two to five second time misalignment was added to determine robustness to a common time distortion.

Two different types of system accuracy estimates were used. The first type assumed that the test song existed in the database. If the closest song in the database was identical to the test song and the similarity value exceeded the predetermined threshold, then the result was considered accurate. Table 2 shows the performance for this accuracy with various time misalignments and distortions. The accuracy decreased slightly when the time misalignment is increased to two seconds. The difference between performance for a one second time misalignment of the design set in Table 1 and the validation set in Table 2 was not statistically significant.

Table 2: Performance of the normalized modulation frequency features when the test song was assumed to exist in the database.

Time Shift	Distortions		
	Compression	*Equalization*	*Normalization*
1 second	99.6 %	97.8 %	99.5 %
2 seconds	99.2 %	96.5 %	99.1 %
random 2-5 seconds	89.6 %	86.2 %	90.5 %

The second type of the system accuracy was based upon the assumption that the unknown test song was not in the database. For this case, a correct result was defined to be no detection when the similarity to all songs in the database was below the predetermined threshold. Table 3 summarizes these results.

Table 3: Performance of the normalized modulation frequency features when the test song was assumed to not be in the database.

Time Shift	Distortions		
	Compression	*Equalization*	*Normalization*
1 second	95.7 %	96.7 %	96.2 %
2 seconds	95.7 %	96.7 %	96.3 %
random 2-5 seconds	96.1 %	96.8 %	96.5 %

5. CONCLUSIONS

This paper discussed joint acoustic and modulation frequency features. A wavelet transform, as a second transform after acoustic frequency detection, was proposed for modulation frequency decomposition. The proposed features were tested in an audio fingerprinting application. The experimental results show that: 1) spectral estimates only were inadequate for this task, 2) modulation frequency features offered substantially improved performance, and 3) subband normalization was necessary for invariance to common time and frequency distortions. These joint frequency features can potentially be applied to other acoustic classification or detection applications such as speaker identification or verification.

Acknowledgement

The authors would like to thank Dr. James W. Pitton for algorithm discussions, John F. Keane and Wibool Piyawattanametha for improving the manuscript, and their friends for helping to create the database.

6. REFERENCES

[1] RIAA and IFPI, Request for Information on Audio Fingerprinting Technologies, 2001

[2] E. Allamanche, J. Herre, O. Hellmuth, et al., " AudioID: Towards Content-Based Identification of Audio Material," 110th AES Convention, Amsterdam, 2001

[3] W. Gardner, " Exploitation of Spectral Redundancy in Cyclostationary Signals," *IEEE Signal Proc. Magazine*, April 1991, pp. 14-36

[4] M. Vinton and L. Atlas, " A Scalable and Progressive Audio Codec," *Proc. of ICASSP*, 3277-80, 2001

[5] L. Cohen, *Time-Frequency Analysis*, Prentice Hall, Englewood Cliffs, NJ, 1995

[6] S. Ewert and T. Dau, " Characterizing Frequency Selectivity for Envelope Fluctuations," *Journal of the Acoustical Society of America*, **108** 1181-96, 2001